Point-MSD: Jointly Mamba Self-Supervised Self-Distilling Point Cloud Representation Learning

Linzhi Su[1], Mengna Yang[1], Jie Liu[2], Xingxing Hao[1,*], Chenyang Zhang[1], Xin Cao[1]
[1]*School of Information Science and Technology, Northwest University, Shaanxi, China*
[2]*College of Computer and Information Engineering, Henan Normal University, Henan, China*
[*]Corresponding Author: Xingxing Hao
Email: sulinzhi029@163.com, yangmengna@stumail.nwu.edu.cn, jieliu2017@126.com, xingxing.hao@nwu.edu.cn, 2634588577@qq.com, caoxin918@hotmail.com.

2024 IEEE 17th International Conference on Signal Processing (ICSP) | 979-8-3503-8738-4/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICSP62129.2024.10846527

***Abstract*—Self-supervised tasks, which extract supervisory signals from the data without human labeling, have garnered significant attention. Mask modeling, in particular, has piqued interest. To leverage unlabeled point data in self-supervised learning, we introduce a self-distilling architecture that integrates mask modeling to enhance the model's ability to extract representations and strengthen its high-level semantic feature learning on two versions of a point cloud. Recognizing that point coordinate embeddings can lead to information disclosure distinct from images and text, we incorporate mask embedding into the decoder of an encoder-decoder structure to mitigate this issue and ensure the model's learning is not overly simplistic. These backbones are typically based on CNNs or Transformers, which have limitations in local receptive field and global modeling, respectively. Drawing inspiration from Mamba, which originates from a state space model with global modeling capabilities and linear complexity, we propose a novel approach called Jointly Mamba Self-Supervised Self-Distilling Point Cloud Representation Learning (Point-MSD). Additionally, we introduce a global semantic consistency loss to enrich the overall shape understanding of 3D objects. Extensive experiments demonstrate that our method can learn superior point cloud representations. Notably, Point-MSD achieves a 93.52% accuracy on ModelNet40, outperforming PointMamba by 0.93%.**

Keywords—point cloud, self-supervised, self-distilling, mamba, mask

I. INTRODUCTION

Point clouds, a crucial data structure, play a vital role in understanding the 3D world. The availability of affordable consumer-grade sensors has made point clouds a richer and more precise source of 3D geometric information for applications such as autonomous driving [1] and robotics [2]. Supervised learning on point clouds has shown promising results [3]. However, most point cloud data are unlabeled, and manual labeling is often time-consuming and laborious. Self-supervised learning, which learns from the data itself without labels, has distinct advantages in fields like natural language processing (NLP) [4], computer vision [5], and 3D point clouds [6]. The primary frameworks for self-supervised learning include contrastive learning [7] and generative learning [8]. Contrastive learning emphasizes feature consistency across different augmented versions of the same sample but overlooks local detail recovery. Generative learning focuses on low-level semantic recovery, neglecting high-level understanding. To address these limitations, we introduce a self-distillation architecture for self-supervised tasks. This architecture not only captures the consistency of high-level semantics within the same sample but also reconstructs high-quality mask information through self-distillation between two branches.

Mask modeling has become a focus in self-supervised tasks for point clouds [9]. Pioneering in this field are Point-MAE [6] and Point-BERT [10], which draw on MAE and BERT methodologies, respectively. Point-MAE has notably gained attention due to its high mask ratio and asymmetric architecture, which eliminates the need for an offline tokenizer and mitigates the issue of location information leakage in point clouds. Building on these insights, we propose an asymmetric architecture that inputs only the visible view into the encoder, enhancing model training difficulty and reducing the encoder's computational load.

The Transformer, a common backbone for self-supervised learning, is favored in NLP [11], computer vision [12], and 3D point clouds [13] due to its attention mechanism that effectively captures long-distance dependencies. However, the Transformer's self-attention mechanism leads to a time complexity of $O(N^2)$, where N is the sequence length. Efforts to address this issue [14] risk losing data's intrinsic structural details, impacting model learning. Convolutional Neural Networks (CNNs) offer linear complexity and encode local information directly, but their limited global understanding restricts their learning capabilities [15]. Drawing inspiration from the Mamba architecture [16], we introduce a model that merges the benefits of Transformers and CNNs. This model features linear complexity and a global attention mechanism, avoiding the $O(N^2)$, limitation and boasting a compact parameter set, with performance that even surpasses that of the Transformer.

Based on the above analysis, we introduce Point-MSD, a self-supervised learning framework that integrates mask mamba with an asymmetric architecture and self-distillation. This architecture reduces encoder input pressure, leverages the benefits of linear complexity and long-distance dependency, and enhances feature extraction. By optimizing both high-level semantics and global shape understanding, our model captures rich 3D shape representations. Extensive experiments on tasks like object classification and semantic segmentation demonstrate that Point-MSD delivers competitive performance compared to existing transformer-based networks.

In summary, our contributions are as follows:

- We introduce Point-MSD, a self-supervised learning method that combines mask mamba with an asymmetric architecture and self-distillation, to

harness the potential of point clouds in self-supervised tasks.

- We utilize Mamba as the backbone, which offers linear complexity and a global attention mechanism, allowing for effective global modeling without incurring square-order complexity.
- We implement an asymmetric architecture in mask modeling tasks, enhancing the model's learning capabilities and aligning it more closely with downstream task requirements, leading to superior performance.
- We conduct extensive experiments on downstream tasks, proving the effectiveness of Point-MSD and achieving competitive results against other transformer-based networks.

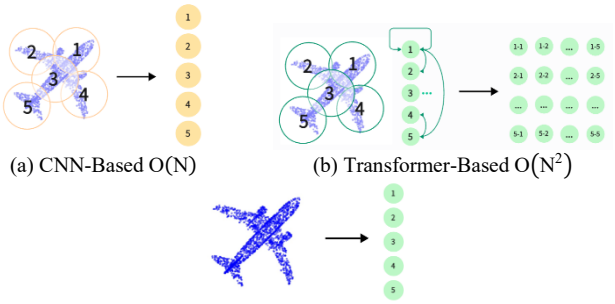
II. Related Work

A. Transformer-based self-supervised learning on point clouds

Transformers have achieved significant success in NLP [4] and computer vision [17], thanks to their attention mechanism that captures global dependencies effectively. The development of transformer-based models has also accelerated in the field of 3D point clouds [18, 19]. However, much of the real-world data is unlabeled, making manual annotation impractical and potentially reducing the effectiveness of supervised learning on point clouds. In this context, self-supervised learning, which extracts supervision signals from the data itself, has become increasingly popular. The Masked Point Modeling (MPM) task is a widely used approach in this area. For example, Point-BERT [10] uses an offline tokenizer to predict masked tokens, P2C [18] leverages prior knowledge of different objects to reconstruct original points, and Joint-MAE [19] integrates 2D information to reconstruct masked information in both 2D images and 3D point clouds. Point-MAE [6] reconstructs masked points under high mask ratios and with asymmetric structure designs. Considering the issue of position information leakage in point clouds and the advantages of MPM tasks, we introduce an asymmetric design to improve model learning efficiency and reduce input pressure to some extent.

B. Mamba

State space models (SSM) have gained popularity for their sequence modeling capabilities. The Structured State-space Sequence model (S4) [20] pioneered the integration of linear SSM into deep learning, showing strong adaptability in handling long sequence data. Building upon this, S5 [21] introduced multiple-input multiple-output SSM and efficient parallel scanning. GSS [22] further enhanced performance with gated state-space layers. To bridge the gap between SSM and transformer attention in language modeling, Fu et al. introduced a new layer called H3 [23]. Subsequently, Mamba [24], a data-dependent SSM layer, was proposed, offering a general language model framework with a selection mechanism and efficient hardware design. In NLP, Mamba models have surpassed traditional transformers, demonstrating linear scalability with input length. As depicted in Fig. 1, CNN-based models offer linear complexity but are biased towards local features, while Transformer-based models have global modeling capabilities but with quadratic complexity. Mamba-based models, in contrast, provide global attention while maintaining linear complexity. The Mamba architecture has been successfully applied in various domains, including image classification [25], multi-modal semantic segmentation [16], and medical image segmentation [26]. However, there has been limited work on Mamba-based self-supervised learning for 3D point clouds [27]. Therefore, we aim to introduce Mamba to self-supervised tasks to explore its potential in this area.



(a) CNN-Based O(N) (b) Transformer-Based $O(N^2)$

(c) Mamba-Based O(N)

Fig. 1. Comparison of complexity on different point cloud process.

C. Self-distillation

In recent times, deep learning networks with millions of parameters have become prevalent. Knowledge distillation, a form of model compression, is widely recognized for its use of teacher-student branches to transfer knowledge from the teacher to guide the student. Typically, the teacher branch has greater network capacity and a more complex structure than the student branch. Self-distillation, a subset of this approach, has gained popularity as it does not require training large models for the teacher branch. Here, the teacher and student branches share the same architecture but have different parameters; the teacher's parameters are usually the moving averages of the student's parameters. This concept has spurred research in computer vision [28] and 3D point clouds [29]. Methods like BYOL [30] integrate self-distillation into feature contrast frameworks, aiming to align the student's features closely with those of the teacher. DINO [31] employs self-distillation to optimize the categorical distribution across teacher-student branches, thereby enhancing learning capabilities. Diverging from these methods, we introduce a Mamba-Based backbone, which combines global attention and linear complexity, to fully leverage the benefits of self-distillation architecture in self-supervised tasks for point clouds.

III. Methods

This study introduces Mamba to point cloud self-supervised mask modeling tasks. We begin with a detailed explanation of Mamba's fundamentals. We then present our framework, which integrates mask mamba with an asymmetric design and self-distillation. To strengthen the model's knowledge acquisition from unlabeled data, we employ a self-distillation architecture that aligns global semantic information between teacher and student branches. To address point cloud position information leakage, we incorporate an asymmetric design in the student branch, enhancing training difficulty and reducing input load. We also leverage the Mamba backbone for its linear complexity and

strong global modeling capabilities. Finally, by combining objective losses from different perspectives on high-level semantics and global shape, we achieve richer point cloud representations suitable for various downstream tasks. The framework is illustrated in Fig. 2.

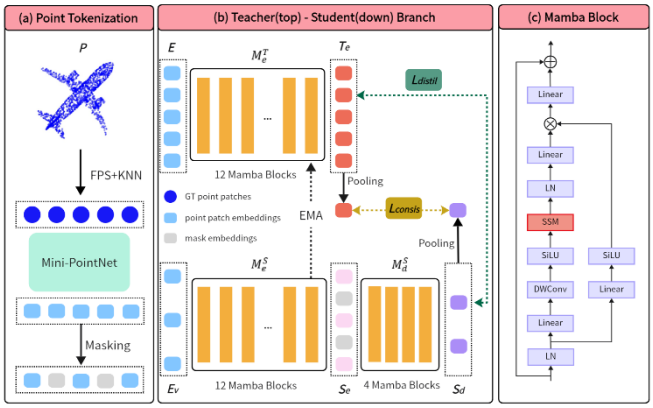


Fig. 2. The framework of our Point-MSD.

A. Mamba

An SSM model with a selection mechanism and efficient hardware design in NLP. SSM is a class of sequence-to-sequence modeling system. Inspired by the continuous linear time-invariant system, which can effectively capture the inherent dynamic of systems through implicit mapping input $x(t)\in R$ to output $y(t)\in R$. Theoretically, this system can be formulated as a linear ordinary differential equation (ODE):

$$\begin{aligned} h'(t)&=Ah(t)+Bx(t),\\ y(t)&=Ch(t)+Dx(t). \end{aligned} \tag{1}$$

where $h(t)\in R^N$ denotes hidden state, and $h'(t)$ refer to the time derivative of $h(t)$. N is the state size, $A\in R^{N\times N}$, $B\in R^{N\times 1}$, $C\in R^{1\times N}$, $D\in R$ are the system matrices parameters.

Since the continuous system is a continuous-time model, only through the discretization process can be integrated into deep learning algorithms. To process discrete sequences like image and text, the commonly used method for discretization is the zero-order hold (ZOH) rule. Concretely, we use timescale parameter Δ to transform the continuous parameters A, B to discrete parameters $\overline{A}, \overline{B}$, which is formulated as follows:

$$\begin{aligned} \overline{A}&=\exp(\Delta A),\\ \overline{B}&=(\Delta A)^{-1}(\exp(\Delta A)-I)\cdot\Delta B,\\ \overline{C}&=C. \end{aligned} \tag{2}$$

After discretization, Eq. (1) can be rewritten as:

$$\begin{aligned} h_k&=\overline{A}h_{k-1}+\overline{B}x_k,\\ y_k&=\overline{C}h_k+\overline{D}x_k. \end{aligned} \tag{3}$$

where $\overline{D}$ typically serving as a residual connection, is often discarded in the context.

Due to the characteristics of linear time-invariant (LTI) SSMs, invariant parameters are obtained regardless of the differences in the input. To address this limitation, we introduce Mamba, which incorporates a selective scan mechanism, making SSM data-dependent and enabling the model to be contextually aware of the input. Using this selection mechanism, Mamba can generate more accurate and efficient representation of the input sequence.

B. Point Tokenization

Given a point cloud $P=\{p_1,p_2,\ldots,p_N\}$, we first sample s center points $C\in R^{s\times 3}$ using the farthest point sampling (FPS). For each center point $c_i\in C$, we select k neighbor points from around and form corresponding a point patch $u_i=\{p_{i1},p_{i2},\ldots,p_{ik}\}\in U$ by using the k-nearest neighbors (KNN) algorithm. These point patches U are then normalized by subtracting the corresponding center point coordinate to obtain relative coordinates. To aggregate point patches local information, mini-PointNet is introduced, which contains a shared MLP and a Max-Pooling layer to obtain point embeddings $E\in R^{s\times C}$. Before entering student branch, we use common random mask strategy with mask radio r to obtain visual embeddings $E_v\in R^{(1-r)s\times C}$.

$$\begin{aligned} U,\ C&=\mathrm{KNN}(\mathrm{FPS}(P)),\\ E&=\mathrm{PointNet}(U),\\ E_v&=\mathrm{Masking}(E). \end{aligned} \tag{4}$$

C. Student Branch

Including a mamba encode $M_e^S=\mathrm{Concat}[M_1,M_2,\ldots,M_{12}]$ and decoder $M_d^S=\mathrm{Concat}[M_1,M_2,M_3,M_4]$, which consist of 12 Mamba Blocks and 4 Mamba Blocks, respectively. For the mamba encoder M_e^S, we put E_v into encoder to generate encoder features $S_e\in R^{(1-r)s\times C}$. To mitigate point cloud position information leakage, we only introduce mask embeddings $E_m\in R^{rs\times C}$ in the decoder, which are obtained using two-layer MLP maps at each center point coordinate. For the mamba decoder M_d^S, we combine mask embedding E_m with encoder features S_e as input, and then output predicted mask features $S_d\in R^{rs\times C}$.

$$\begin{aligned} S_e&=M_e^S(E_v),\\ S_d&=M_d^S(S_e,E_m). \end{aligned} \tag{5}$$

D. Teacher Branch

A simple mamba encoder $M_e^T=\mathrm{Concat}[M_1,M_2,\ldots,M_{12}]$ consisting of 12 Mamba Blocks. Our mamba encoder takes E as input and outputs GT masked features $T_e\in R^{rs\times C}$ corresponding to the student branch predicted masked features S_d. Notably, the student network parameters θ^s are exponentially moving averaged (EMA) to the teacher network parameters θ^t. Specifically, θ^t is updated according to $\theta^t\leftarrow\alpha\theta^t+(1-\alpha)\theta^s$, where $\alpha\in[0,1)$ is the EMA decay rate.

$$T_e=M_e^T(E). \tag{6}$$

E. Mamba Block

A mamba block structure is shown in Fig. 2(c), which consists of LayerNormal (LN), Linear projection layer (Linear), SSM, depth-wise convolution (DWConv), SiLU [32], and residual connections. For a Mamba Block M_i, we first execute deep convolution processing to obtain M_i^1 after normalization and Linear projection layer, and then apply SiLU. Secondly, to improve the performance of the model when processing discrete data, we use SSM module to update the state and calculate the output result. Thirdly, we conduct residual connection to combine the output after SiLU and Linear operations with the input origin input M_{i-1} after LN, Linear, and SiLU operations, which makes the model learning more flexible and maintain the identity mapping of input and output. Finally, we use Linear combined with M_{i-1} to obtain final M_i. The overall process can be formulated as follows:

$$\begin{aligned} M_i^1 &= \mathrm{DWConv}\left(\mathrm{Linear}\left(\mathrm{LN}(M_{i-1})\right)\right), \\ M_i^2 &= \mathrm{Linear}\left(\mathrm{SSM}\left(\mathrm{SiLU}\left(M_i^1\right)\right)\right) \times \mathrm{SiLU}\left(\mathrm{Linear}\left(\mathrm{LN}(M_{i-1})\right)\right), \\ M_i &= \mathrm{Linear}\left(M_i^2\right) + M_{i-1}. \end{aligned} \tag{7}$$

where M_i represents i-th mamba block output and $M_0=E$.

F. Local Distillation Loss

In addition to the low-level semantics reconstruction, the high-level semantics feature contrast is also crucial for model learning. Thus, we use Smooth L1 Loss as mask features optimization function:

$$L_{distil} = \frac{1}{rs}\begin{cases} 0.5\left(T_e^i - S_d^i\right)^2, & \text{if} \left|T_e^i - S_d^i\right| < 1 \\ \left|T_e^i - S_d^i\right| - 0.5, & \text{otherwise} \end{cases} \tag{8}$$

where T_e^i is the i-th target value of the teacher branch T_e, S_d^i is the i-th predicted value of the student branch S_d.

G. Global Semantic Consistency Loss

Considering optimization only from the perspective of local features, this strategy may affect global semantic learning, so we introduce global feature loss by using Smooth L1 Loss, to learn global semantic information.

$$L_{consis} = \begin{cases} 0.5(\overline{T}_e - \overline{S}_d)^2, & \text{if} \left|\overline{T}_e - \overline{S}_d\right| < 1 \\ \left|\overline{T}_e - \overline{S}_d\right| - 0.5, & \text{otherwise} \end{cases} \tag{9}$$

where $\overline{T}_e$ is the average of target values of the teacher branch T_e, and $\overline{T}_e = \frac{1}{s}\sum_{i=1}^{s} \overline{T}_e^i$. S_d is the average of predicted values of the student branch S_d, and $\overline{S}_d = \frac{1}{rs}\sum_{i=1}^{rs} \overline{S}_d^i$.

The overall loss function is defined as:

$$L_{all} = L_{distil} + L_{consis} \tag{10}$$

IV. Experiments

In this section, we detail the self-supervised pre-training setup of our model on the ShapeNet [33] dataset. We then assess the model's performance on object classification and semantic segmentation tasks. Lastly, we conduct an in-depth analysis of the model's learning effectiveness across various module configurations.

A. Dataset

Following existing studies to pre-train our model Point-MSD on ShapeNet, which contains 57448 3D instances from 55 object categories. During pre-training, we sample 1024 points from each instance and conduct data augmentations by randomly scaling and translation. Then, we partition it into 64 patches, where each patch contains 32 points.

B. Training setups

For the encoder, we use 12 Mamba Blocks with the embedding of 384 and 6 heads. To form a lightweight decoder, the depth of the decoder is set to 4, and the number of heads is set to 6. We pre-train our Point-MSD for 800 epochs with a batch size of 128. To equal PointMamba and our model, we employ the same mask strategy to maintain the consistency of encoder input. We employ the AdamW [34] optimizer and a cosine learning rate scheduler [35], and initial weight decay and learning rate are set to be 0.001 and 0.05. Referring to data2vec [36], we set $\beta=2$ as the parameter for Smooth L1 Loss.

C. Downstream tasks

We evaluate our pre-trained model on ModelNet40 [37], which contains over 10000 objects from 40 categories. During the fine-tuning process, we sample 1024 points as input and use scaling, centering, and unit sphering as data augmentations. The classification header is set to 3-layer MLP, which handle the features of Mamba encoder pooling. The experimental results are shown in Table 1. Comparing with current transformer-based methods, our model achieves competitive results. On the one hand, the model increases Transformer+OcCo, Point-BERT, and MaskSurf by 1.42%, 0.32%, 0.12%, it shows that ability of representation learning on Mamba. On the other hand, the model decreased Point-MAE, Joint-MAE, and Point-RAE by 0.28%, 0.48%, 0.58%, it may be Mamba is missing channel-wise attention and not fully utilize the advantage of global attention. In the case of both two branch architectures, the Point-RAE based on Transformer has nearly 3× as many parameters as our Point-MSD. Comparing with existing Mamba-Based studies on 3D Point Clouds (PointMamba, PCM), we can find that our model achieves the increase PointMamba, PCM by 0.93%, 0.12%, which demonstrates knowledge learned is more robust after introducing self-distillation architecture.

TABLE I. Object classification on ModelNet40 ※ denotes reproduced results

Methods	Params (M)	Accuracy (%)
Transformer+OcCo [40]	6.86	92.10
Point-BERT [10]	22.10	93.20
MaskSurf [41]	29.04	93.40

Methods	Params (M)	Accuracy (%)
3D-OAE [42]	23.23	93.40
Point-MAE [6]	22.10	93.80
Joint-MAE [19]	-	94.00
Point-RAE [44]	162.63	94.10
PointMamba [27]※	12.30	92.59
PCM[43]	34.20	93.40
Point-MSD (Ours)	55.50	93.52

We evaluate transfer ability of model representations learning on ScanObjectNN [38], which contains 2902 instances from 15 categories and consists of three variants: OBJ-BG, OBJ-ONLY, and PB-T50-RS. In PB-T50-RS setting, we found our Point-MSD exceeds the baseline by 6.43%. Comparing other transformer-based methods, such as Point-Bert, 3D-OAE, and Point-MAE, our model gains 2.15%, 2.05%, and 0.04%, respectively. It shows that learned representation of our model can transfer to the real-world application domain. However, mamba lack channel-wise interaction in global attention modeling, which can influence the ability of our model extracts feature representation. Thus, Contrasting to SOTA method (Point-RAE), our model decreases by 2.28%. From a mamba-based perspective, our model increases 0.35% compared to PointMamba, it indicates that the introduction of self-distillation architecture is conducive to more generalized learning of the model. Comparing with PCM, which combines point order and mamba, exceeds our model by 2.88%, it learns rich point information passed by inputting points from multi-angle and compensates channel-wise interaction for mamba to a certain extent.

TABLE II. OBJECT CLASSIFICATION ON SCANOBJECTNN (%)

Methods	OBJ-BG	OBJ-ONLY	PB-T50-RS
Transformer+OcCo [40]	84.85	85.54	78.79
Point-BERT [10]	87.43	88.12	83.07
MaskSurf [41]	91.22	89.17	85.81
3D-OAE [42]	89.16	88.64	83.17
Point-MAE [6]	90.02	88.29	85.18
Joint-MAE [19]	90.94	88.86	86.07
Point-RAE [44]	91.20	90.40	87.50
PointMamba [27]	90.71	88.47	84.87
PCM [43]	-	-	88.10
Point-MSD (Ours)	89.67	88.81	85.22

We transfer our model to 3D indoor scene semantic segmentation of large-scale scenes [39], which provides 6 different indoor areas that consist of 272 rooms from 13 semantic classes. Following STRL settings, we use 4096 points with only geometric features (x, y, z) as input and train model for 100 epochs with batch size 24. We report the experimental results in Table 3. Compare with PointMamba, our model demonstrates better transfer ability through introduce self-distillation learning mode. The results show that the overall accuracy and mean intersection of union are increased by 3.86% and 13.93%, respectively.

TABLE III. SEMANTIC SEGMENTATION ON S3DIS. OVERALL ACCURACY (OA) (%) AND MEAN INTERSECTION OF UNION (MIOU) (%) ON THE S3DIS ACROSS SIX FOLDS ※ DENOTES REPRODUCED RESULTS

Methods	OA	mIoU
PointMamba※	75.56	34.41
Point-MSD (Ours)	79.42	48.34

Comparing with traditional self-distillation architecture, we introduce mask strategy and encoder-decoder structure in student branch. To explore the influence of mask embeddings on different positions (encoder and decoder), we conduct comparative trials. As shown in Table 4, the mask embedding place in decoder instead of encoder, which gains 0.41% on ModelNet40 and 1.94% on the PB-T50-RS setting of ScanObjectNN. It demonstrates that the point clouds easily produce position information leakage by using position coordinates, which differ from image and text by index. Thus, Adding the mask into the decoder is better for the model to learn robust representation.

TABLE IV. COMPARISON RESULTS UNDER DIFFERENT MASK EMBEDDING POSITIONS

Methods	Params (M)	ModelNet40 (%)	ScanObjectNN (%)		
			OBJ-BG	*OBJ-ONLY*	*PB-T50-RS*
Encoder	48.1	93.11	87.95	87.78	83.28
Decoder (Ours)	55.5	93.52	89.67	88.81	85.22

To obtain global semantic understanding of 3D objects, we conduct different average pooling operation on features of mamba blocks. As shown in Table 5, the last 6 mamba blocks can learn rich global semantic information and achieve 93.52% accuracy. The former 6 mamba blocks attain suboptimal result with 93.48%, which may be the first mamba block is global and the later blocks gradually transition from local to global. Thus, comparing to average later mamba blocks, merging all mamba blocks makes accuracy decreasing 0.37%.

TABLE V. AVERAGE POOLING RESULTS ON DIFFERENT MAMBA BLOCKS (%)

Mamba Block Index	Accuracy
[0,1,2,3,4,5,6,7,8,9,10,11]	93.15
[0,1,2,3,4,5]	93.48
[6,7,8,9,10,11]	93.52

TABLE VI. CLASSIFICATION RESULTS UNDER DIFFERENT LOSS ITEMS (%)

Loss Items	Accuracy
Base	92.79
Base+ L_{consis}	93.52

In model learning process, global shape semantic consistency is important for understanding of 3D objects. During pre-training, global semantic implicitly utilize mask

view and origin view of a point cloud as positive pair due to their same optimal objective, and then reduce the distance in feature space. As shown in Table 6, compared with base distillation loss (Base), our model achieves increasing by 0.73% after adding global semantic loss.

To explore a proper mask strategy, our model compares two different mask methods (random and block) with different mask radios (45%, 65%, and 85%). The object classification results as shown in Table 7. Comparing with other two mask radios, the 65% mask radio achieves higher accuracy whether in random or block mask strategies. It is because of the high redundancy of point cloud data that too much information is retained make model training is too simple, which hinders the learning ability of the model. Furthermore, retaining fewer points also affect the model's ability to acquire the original object. Besides, we can find that random strategy obtains better results instead of block strategy, Thus, we use random strategy with 65% mask radio as our model's experimental settings.

TABLE VII. CLASSIFICATION RESULTS UNDER DIFFERENT MASK STRATEGY WITH DIFFERENT MASK RADIO (%)

Mask strategy	Mask radio	Accuracy
random	45%	92.87
	65%	93.52
	85%	93.31
block	45%	92.87
	65%	93.23
	85%	93.31

The design of backbone is crucial for model representation learning. To study the influence of backbone network on model learning robust representation during training and testing, we conduct experimental comparison of object classification on different backbones, including Transformer and Mamba. The experimental results are shown in Table 8. Compared to Mamba backbone, using Transformer as the backbone improves model accuracy by 0.32%, but the parameters are approximately increased by 2 times. It demonstrates the powerful potential of mamba in self-supervised tasks.

TABLE VIII. CLASSIFICATION RESULTS UNDER DIFFERENT BACKBONE

Backbone	Params (M)	Accuracy (%)
Transformer	100.8	93.84
Mamba	55.5	93.52

V. CONCLUSION

In this paper, we introduce a method, Jointly Mamba Self-Supervised Self-Distilling Point Cloud Representation Learning. Our approach utilizes the Mamba backbone as an effective alternative to the Transformer, offering global attention modeling with linear complexity that is readily applicable to point clouds without additional operations.

To enhance the model's ability to extract representations from unlabeled data, we integrate a self-distilling learning paradigm featuring a teacher-student branch. This setup trains the encoder to better capture representations that can be transferred to downstream tasks. Furthermore, we introduce a global semantic consistency loss to deepen the model's understanding of overall shape semantics.

Extensive experiments show that our Point-MSD achieves competitive results in object classification and semantic segmentation tasks. However, while our method generalizes well to other downstream tasks, it does not significantly outperform state-of-the-art (SOTA) methods in object classification. This could be due to Mamba's data-dependent nature and the lack of channel-wise interaction, potentially impacting the robustness of feature learning.

In the future, we will explore the issue of channel-wise interaction learning in 3D point clouds to improve robustness. We believe our model offers valuable insights for further exploration of Mamba in self-supervised tasks.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (61806164, 62106199, 61701403), National Major Scientific Research Instrument Development Projects of China (82127805), Key Research and Development Program of Shaanxi Province (2024SF-YBXM-681, 2019GY215, 2021ZDLSF06-04).

REFERENCES

[1] Q. H. Pham, P. Sevestre, R. S. Pahwa, H. J. Zhan, C. H. Pang, Y. D. Chen, A. Mustafa, V. Chandrasekhar and J. Lin, "A* 3d dataset: Towards autonomous driving in challenging environments," in *2020 IEEE International conference on Robotics and Automation (ICRA)*, IEEE, 2020.

[2] L. Landrieu and M. Simonovsky, "Large-scale point cloud semantic segmentation with superpoint graphs," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4558-4567, 2018.

[3] G. C. Qian, Y. C. Li, H. W. Peng, J. J. Mai, H. Hammoud, M. Elhoseiny and B. Ghanem, "Pointnext: Revisiting pointnet++ with improved training and scaling strategies," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23192-23204, 2022.

[4] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] K. M. He, X. L. Chen, S. N. Xie, Y. H. Li, P. Dollar and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000-16009, 2022.

[6] Y. T. Pang, W. X. Wang, F. E. Tay, W. Liu, Y. H. Tian and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," in *European conference on computer vision*, pp. 604-621, Springer, 2022.

[7] S. Y. Huang, Y. C. Xie, S. C. Zhu and Y. X. Zhu, "Spatio-temporal self-supervised representation learning for 3d point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.6535-6545, 2021.

[8] X. Y. Tian, H. X. Ran, Y. Wang and H. Zhao, "Geomae: Masked geometric target prediction for self-supervised point cloud pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.13570-13580, 2023.

[9] Y. Tang, X. Z. Li, J. F. Xu, Q. Yu, L. Hu, Y. X. Hao and M. Chen, "Point-LGMask: Local and Global Contexts Embedding for Point Cloud Pre-training with Multi-Ratio Masking," *IEEE Transactions on Multimedia*, IEEE, 2023.

[10] X. M. Yu, L. L. Tang, Y. M. Rao, T. J. Huang, J. Zhou and J. W. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.19313-19322, 2022.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever and others, "Language Models are Unsupervised Multitask Learner," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[12] S. Hirose, N. Wada, J. Katto and H. M. Sun, "Vit-gan: Using vision transformer as discriminator with adaptive data augmentation," in *2021 3rd International Conference on Computer Communication and the Internet (ICCCI)*, pp. 185-189, IEEE, 2021.

[13] K. X. Fu, M. Z. Yuan, S. L. Liu and M. N. Wang, "Boosting Point-BERT by Multi-choice Tokens," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 438-447, IEEE, 2023.

[14] D. C. Han, X. R. Pan, Y. Z. Han, S. J. Song and G. Huang, "Flatten transformer: Vision transformer using focused linear attention," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5961-5967, 2023.

[15] H. Thomas, C. R. Qi, J. E. Deschaud, B. Marcotegui, F. Goulette and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6411-6420, 2019.

[16] Z. F. Wan, Y. H. Wang, S. L. Yong, P. P. Zhang, S. Stepputtis, K. Sycara and Y. Q. Xie, "Sigma: Siamese Mamba Network for Multi-Modal Semantic Segmentation," *arXiv preprint arXiv:2404.04256*, 2024.

[17] H. B. Bao, L. Dong, S. H. Piao and F. R. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[18] R. K. Cui, S. Qiu, S. Anwar, J. W. Liu, C. Y. Xing, J. Zhang and N. Barnes, "P2c: Self-supervised point cloud completion from single partial clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14351-14360, 2023.

[19] Z. Y. Guo, R. R. Zhang, L. T. Qiu, X. Z. Li and P. A. Heng, "Joint-MAE: 2D-3D joint masked autoencoders for 3D point cloud pre-training," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023.

[20] A. Gu, K. Goel and C. Ré, "Efficiently Modeling Long Sequences with Structured State Spaces," in *International Conference on Learning Representations*, 2021.

[21] J. T. Smith, A. Warrington and S.W. Linderman, "Simplified State Space Layers for Sequence Modeling," in *The Eleventh International Conference on Learning Representations*, 2022.

[22] H. Mehta, A. Gupta, A. Cutkosky and B. Neyshabur, "Long Range Language Modeling via Gated State Spaces," in *International Conference on Learning Representations*, 2023.

[23] D. Y. Fu, T. Dao, K. K. Saab, A. W. Thomas, A. Rudra and C. Re, "Hungry Hungry Hippos: Towards Language Modeling with State Space Models," in *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

[24] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[25] Y. Liu, Y. J. Tian, Y. Z. Zhao, H. T. Yu, L. X. Xie, Y. W. Wang, Q. X. Ye and Y. F. Liu, "Vmamba: Visual state space model," *arXiv preprint arXiv:2401.10166*, 2024.

[26] J. Ma, F. Li and B. Wang, "U-mamba: Enhancing long-range dependency for biomedical image segmentation," *arXiv preprint arXiv:2401.04722*, 2024.

[27] D. K. Liang, X. Zhou, X. Y. Wang, X. K. Zhu, W. Xu, Z. K. Zou, X. Q. Ye and X. Bai, "PointMamba: A Simple State Space Model for Point Cloud Analysis," *arXiv preprint arXiv:2402.10739*, 2024.

[28] J. Jang, S. Kim, K. Y. Yoo, C. Kong, J. Kim and N. Kwak, "Self-distilled self-supervised representation learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

[29] T. Furuya, Z. Chen, R. Ohbuchi and Z. Z. Kuang, "Self-supervised learning of rotation-invariant 3D point set features using transformer and its self-distillation," *Computer Vision and Image Understanding*, vol. 244, p. 104025, Elsevier, 2024.

[30] J.-B. Grill, F. Strub, F. Altche, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, P. B. Avila, Z. H. Guo, A. M. Gheshlaghi and other, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21271-21284, 2020.

[31] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650-9660, 2021.

[32] N. Shazeer, "GLU Variants Improve Transformer," *arXiv preprint arXiv:2002.05202*, 2020.

[33] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. X. Huang, Z. M. Li, S. Savarese, M. Savva, S. R. Song, H. Su and other, "ShapeNet: An Information-Rich 3D Model Repository," *arXiv preprint arXiv:1512.03012*, 2015.

[34] I. Loshchilov, and F. Hutter, "Decoupled Weight Decay Regularization," in *International Conference on Learning Representations*, 2018.

[35] I. Loshchilov and F. Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts," in *International Conference on Learning Representations*, 2016.

[36] A. Baevski, W. N. Hsu, Q. T. Xu, A. Babu, J. T. Gu and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," in *International Conference on Machine Learning*, pp. 1298-1312, PMLR, 2022.

[37] Z. R. Wu, S. R. Song, A. Khosla, F. Yu, L. G. Zhang, X. O. Tang and J. X. Xiao, "3D ShapeNets: A Deep Representation for Volumetric Shapes," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1912-1920, 2015.

[38] M. A. Uy, Q. H. Pham, B. S. Hua, T. Nguyen and S. K. Yeung, "Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1588-1597, 2019.

[39] I. Armeni, O. Sener, A. Zamir, H. L. Jiang, I. Brilakis, M. Fischer and S. Savarese, "3d semantic parsing of large-scale indoor spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1534-1543, 2016.

[40] H. Wang, Q. Liu, X. Y. Yue, J. Lasenby and M. J. Kusner, "Unsupervised point cloud pre-training via occlusion completion," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9782-9792, 2021.

[41] Y. B. Zhang, J. H. Lin, C. H. He, Y. W. Chen, K. Jia and L. Zhang, "Masked Surfel Prediction for Self-Supervised Point Cloud Learning," *arXiv preprint arXiv:2207.03111*, 2022.

[42] J. S. Zhou, X. Wen, B. R. Ma, Y. S. Liu, Y. Gao, Y. Fang and Z. Z. Han, "3d-oae: Occlusion auto-encoders for self-supervised learning on point clouds," *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 15416-15423, IEEE, 2024.

[43] T. Zhang, X. T. Li, H. B. Yuan, S. P. Ji and S. C. Yan, "Point Could Mamba: Point Cloud Learning via State Space Model," *arXiv preprint arXiv:2403.00762*, 2024.

[44] Y. Liu, C. Chen, C. Wang, X. L. King and M. Y. Liu, "Regress Before Construct: Regress Autoencoder for Point Cloud Self-supervised Learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, pp.1738-1749, 2023.